AN EXAMINATION OF PROCEDURES FOR DETERMINING
THE NUMBER OF CLUSTERS IN A DATA SET

GLENN W. MILLIGAN AND MARTHA C. COOPER
THE OHIO STATE UNIVERSITY

A Monte Carlo evaluation of 30 procedures for determining the number of clusters was conducted on artificial data sets which contained either 2, 3, 4, or 5 distinct nonoverlapping clusters. To provide a variety of clustering solutions, the data sets were analyzed by four hierarchical clustering methods. External criterion measures indicated excellent recovery of the true cluster structure by the methods at the correct hierarchy level. Thus, the clustering present in the data was quite strong. The simulation results for the stopping rules revealed a wide range in their ability to determine the correct number of clusters in the data. Several procedures worked fairly well, whereas others performed rather poorly. Thus, the latter group of rules would appear to have little validity, particularly for data sets containing distinct clusters. Applied researchers are urged to select one or more of the better criteria. However, users are cautioned that the performance of some of the criteria may be data dependent.

Key words: classification, stopping rules, numerical taxonomy.

Introduction

In most real life clustering situations, an applied researcher is faced with the dilemma of selecting the number of clusters or partitions in the final solution (Everitt, 1979; Sneath & Sokal, 1973). Virtually all clustering procedures provide little if any information as to the number of clusters present in the data. Nonhierarchical procedures usually require the user to specify this parameter before any clustering is accomplished and hierarchical methods routinely produce a series of solutions ranging from n clusters to a solution with only one cluster present (assume n objects in the data set). As such, numerous procedures for determining the number of clusters in a data set have been proposed (Dubes & Jain, 1979; Milligan, 1981c; Perruchet, 1983). When applied to the results of hierarchical clustering methods, these techniques are sometimes referred to as stopping rules. Often, such rules can be extended for use with nonhierarchical procedures as well.

The application of a stopping rule in a cluster analytic situation can result in a correct decision or in a decision error. Basically, two different types of decision errors can result. The first kind of error occurs when the stopping rule indicates \( k \) clusters are present when, in fact, there were less than \( k \) clusters in the data. That is, a solution containing too many clusters was obtained. The second kind of error occurs when the stopping rule indicates fewer clusters in the data than are actually present. Hence, a solution with too few clusters was obtained. Although the severity of the two types of errors would change depending on the context of the problem, the second type of error might be considered more serious in most applied analyses because information is lost by merging distinct clusters.

The present study reports the results of a simulation experiment designed to determine the validity of 30 stopping rules already in the clustering literature. Although a vast number of references exist, few comparative studies have been performed on these mea-
sures. Authors continue to introduce new stopping criteria while providing little or no comparative performance information. Since the stopping rules are heuristic, ad hoc procedures, an applied researcher must critically examine the suggested solution provided by any such index. The present simulation results should help applied researchers in this evaluation task.

The remainder of the paper is organized in the traditional method, results, and discussion sections format. In the methods section, a description of the test data sets is provided along with a discussion of the 30 stopping rules. The results section presents the findings of the simulation study. Finally, the discussion section interprets the results, offers recommendations for the application of the stopping rules in applied situations, and gives suggestions for continued research.

Method

Data Sets

The artificial data sets used in the present study contained either 2, 3, 4, or 5 distinct nonoverlapping clusters. The data sets consisted of a total of 50 points each and the clusters were embedded in either a 4, 6, or 8 dimensional Euclidean space. Overlap of cluster boundaries was not permitted on the first dimension of the variable space. The absolute minimum separation between neighboring cluster boundaries on the first dimension was equal to .25 times the sum of the within-cluster standard deviations from the two respective clusters. The actual distribution of the points within clusters followed a (mildly) truncated multivariate normal distribution. Hence, the resulting structure could be considered to consist of "natural" clusters which exhibited the properties of external isolation and internal cohesion.

Since a larger number of dimensions tended to contain more (redundant) information as to the clustering in the data, cluster recovery by the methods tended to increase with increasing dimensionality. Similar results were found for the best stopping rules. That is, the better rules capitalized on this redundant information and thus exhibited greater accuracy. The poorer rules tended to display fairly constant recovery as the number of dimensions increased.

The design factors corresponding to the number of clusters and to the number of dimensions were crossed with each other and both were crossed with a third factor that determined the number of points within the clusters. This third factor consisted of three levels where one level had an equal number of points in each cluster (or as close to equality as possible). The second level required that one cluster must always contain 10% of the data points, whereas the third level required that one cluster must contain 60% of the items. The remaining points were distributed as equally as possible across the other clusters present in the data. The 60% condition produced a marked discrepancy in cluster sizes for data sets with larger number of clusters while the 10% condition produced a discrepancy when few clusters were present. Overall, there were 36 cells in the design. Three replications were generated in each cell of the design. This produced 108 data sets for testing purposes. Each data set was used to compute a dissimilarity matrix consisting of the Euclidean distances between points. The matrix of distances was the input data for the clustering methods. Each matrix was analyzed by the four clustering methods to provide a variety of solutions. Thus, the design produced a total of 432 test solutions. The four hierarchical clustering methods used to generate the solutions were the single link, complete link, group average, and Ward's minimum variance procedures. The data generation process used in the present experiment corresponds to the error-free data conditions in previous studies (Milligan 1980; Milligan, 1981b).

It is useful to stress the fact that the clusters were internally cohesive and well sepa-